AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 23, Number 3, 2007, pp. 477–481

DOI: 10.1089/aid.2006.0232

Sequence Note

HIV Type 1 Subtype C *gag* and *nef* Diversity in Southern Africa

HELBA BREDELL,[1] DARREN P. MARTIN,[1] JOANNE VAN HARMELEN,[1] ARVIND VARSANI,[2] HAYNES W. SHEPPARD,[3] RICHARD DONOVAN,[3] CLIVE M. GRAY,[4] HIVNET028 STUDY TEAM,[5] and CAROLYN WILLIAMSON[1]

ABSTRACT

Several HIV-1 subtype C-specific *gag*- and/or *nef*-based vaccines are currently intended for clinical trial in southern Africa. Here we provide sequences of 64 *gag* and 45 *nef* genes sampled in Malawi, Zambia, Zimbabwe, and South Africa and assess the degree of southern African HIV-1 diversity that will confront these vaccines. Whereas reasonable phylogenetic evidence exists for geographical clustering of subtype C *gag* and *nef* sequences from various other parts of the world, there is little evidence of similar population founder effects in the southern African epidemic. The entire breadth of subtype C diversity is represented in the southern African genes suggesting there may be no advantage in producing region- or country-specific subtype C vaccines. We do not, however, find much evidence of intersubtype recombination in the Southern African genes, implying that the likelihood of vaccine failure due to the emergence of intersubtype recombinants is probably low.

Southern African countries are experiencing the greatest burden of the AIDS epidemic as a result of the devastating spread of human immunodeficiency virus type 1 (HIV-1) subtype C.[1,2] Estimated southern African adult HIV prevalence rates of over 30% in some regions highlight the need for an effective vaccine (http://www.who.int/emc-hiv/fact_sheets/).

One of the most important challenges in vaccine design is dealing with the high degrees of HIV-1 genetic diversity that are a characteristic of the AIDS epidemic. Fueled by high viral mutation, recombination, and replication rates and driven by strong host immune pressures,[3,4] the diversity of global HIV populations is continuing to increase at an astonishing rate. There is, however, considerable variability in the rate at which different regions of the HIV genome are diversifying. Differences in the degree of nucleotide sequence conservation across the HIV genome reflect in part the balance that exists between selection pressures exerted by host immune systems[5] and the functional constraints of viral proteins.[6]

Gag and the central region of Nef are reasonably conserved HIV proteins with high epitope densities that are among the most common targets of host cellular immune responses.[7–9] Conflicting clinical data exist on the association between T cell responses and plasma viremia. While some studies showed no correlation between plasma viral load and $CD8^+$ T cell responses specific against HIV-1 p24-Gag and Nef,[8,10] other studies showed an inverse correlation between HIV-1 C Gag T cell responses and plasma viral load indicating that Gag-specific cytotoxic T lymphocyte (CTL) responses are possibly important in controlling viremia.[9,11] *Gag* and *nef* are therefore attractive potential vaccine components and, accordingly, there are currently several *gag*- and/or *nef*-based vaccines intended for clinical trial in southern Africa (http://www.iavireport.org/trialsdb IAVI database of AIDS vaccines in human trials).

While there is presently a large amount of data available for South African and Botswanan subtype C *gag* and *nef* sequences, the same is not true for other southern African countries. There-

[1]Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, Cape Town, South Africa, 7925.
[2]Electron Microscope Unit, University of Cape Town, Rondebosch, Cape Town, South Africa, 7954.
[3]California Department of Health Sciences, Richmond, California 94804.
[4]AIDS Unit, National Institute for Communicable Diseases, Sandringham, Johannesburg 2131, South Africa.
[5]The HIVNET028 Study group members are listed in the Acknowledgments.

fore, to better assess the degree of HIV genetic diversity confronting proposed southern African subtype C-specific *gag*- and/or *nef*-based vaccines, 64 *gag* and 45 *nef* sequences were determined from 71 HIV-1-infected individuals recruited as part of the HIVNET028 study[9,11] from five clinical sites in southern Africa: Johannesburg ($n = 10$; *gag* = 8; *nef* = 6) and Durban ($n = 22$; *gag* = 22; *nef* = 14) in South Africa; Lusaka in Zambia ($n = 20$; gag = 18; *nef* = 13); Harare in Zimbabwe ($n = 9$; *gag* = 6; *nef* = 5); and Blantyre in Malawi ($n = 10$; *gag* = 10; *nef* = 7).

Viral RNA was extracted from plasma using the QIAamp Viral RNA Kit (QIAGEN, Chatsworth, CA) or the NASBA RNA extractor kit (Life Sciences, Inc.). The *gag* and *nef* genes were reverse transcribed using the ThermoScript RT-PCR System (Invitrogen Corp., San Diego, CA), and an oligo$(dT)_{20}$ primer, and the primer Gag D reverse (5′AATTCCTCCTATCATTTTTGG3′, at position 2382–2402 of the HXB2 genome) and the primer nefOR (5′AGGCAAGCTTTATTGAGG3′, at position 9608–9625 in HXB2), respectively. A 1673-bp full-length *gag* gene product was amplified from cDNA by nested polymerase chain reaction (PCR) using the outer primers Gag D forward (5′TCTCTAGCAGTGGCGCCCG3′, at position 626–644) and Gag D reverse. Three fragments of ~600 bp, overlapping by 77 and 99 nucleotides, respectively, were generated using inner primers: Gag A forward (5′CTCTCGACGCAGGACTCGGCTT3′, 683–704) and Gag A reverse (5′ACATGGGTATCACTTCTGGGCT3′, 1282-1303), Gag B forward (5′CCATATCACCTAGAACTTTGAAT3′, 1226–1248) and Gag B reverse (5′CTCCCTGACATGCTGTCATCAT3′, 1825–1846), and Gag C forward (5′CCTTGTTGGTCCAAAATGCGA 3′, 1748–1768) and Gag C reverse (5′TCTTCTAATACTGTATCATCTGC3′, 2334–2356). A 750-bp full-length *nef* gene product was amplified from cDNA by nested PCR using the outer primers nefOF (5′GTGGAATTCCTGGGACG3′, 8568–8584) and nefOR and inner primers nefIF (5′ CCTAGAAGAATAAGACAGGGC3′, 8754–8774) and nefIR (5′ CTTATATGCAGCATCTGAGG3′, 9498–9517).

PCR products were purified (QIAquick PCR Purification kit,QIAGEN, Chatsworth, CA) and sequenced with the forward and reverse Gag A, B, and C or Nef IF and IR primers using the BigDye Terminator Cycle sequencing kit version 3.1 (Applied Biosystems Inc., Foster City, CA) and an automated ABI PRISM 3100 genetic analyzer (Applied Biosystems Inc.).

Full length *gag* and *nef* sequences were respectively aligned using POA (gap open penalty = 12, gap extension penalty = 4)[12] with all nearly full length *gag* ($n = 1236$) and *nef* ($n = 1923$) sequences available in the Los Alamos National Laboratory HIV-1 sequence database (http://hiv-web.lanl.gov/content/hiv-db) on 4 August 2005 (for *gag*) and 10 February 2006 (for *nef*). Except for subtype and circulating recombinant form (CRF) references and one HIV-1 N sequence, all sequences that did not cluster among the subtype C sequences in a neighbor-joining tree (transition:transversion ratio = 2.0, constructed with PHYLIP v3.5c)[13] were discarded to obtain *gag* and *nef* datasets containing 558 and 677 sequences, respectively. These datasets were realigned using POA and potential inter- and intrasubtype recombinants among the HIVNET028 sequences were identified using six recombination detection methods implemented in RDP3 (default settings for all methods and a Bonferroni corrected *p*-value cutoff = 0.05).[14]

None of the HIVNET028 *gag* or *nef* sequences were identified as being intersubtype recombinants. However, eight HIVNET028 *gag* and three HIVNET028 *nef* sequences were identified by two or more recombination detection methods as being intrasubtype C recombinants. Recombination can seriously impact the validity of phylogenetic inferences[15,16] and we therefore removed the pieces of these HIVNET028 sequences that appeared to have a recombinant origin. Sequences between the approximate recombination breakpoints identified for the 11 detected events were replaced in the alignment with gap characters. In no case was more than 39% of a sequence replaced by gap characters. To further minimize the impact of recombination on subsequent phylogenetic reconstructions, 163 potentially recombinant *gag* sequences and 245 potentially recombinant *nef* sequences, identified by any one of the six recombination detection methods with the same Bonferroni *p*-value correction used to identify HIVNET028 recombinants, were completely removed from the *gag* and *nef* alignments. An additional 14 *gag* and 8 *nef* sequences identified elsewhere as being intersubtype recombinants but that were not identified as such during the recombination screen were also removed from the *gag* and *nef* datasets.

None of the HIVNET028 sequences was more closely related to previously identified intersubtype recombinants with subtype C-like *gag* or *nef* sequences than to other apparently "pure" subtype C sequences. Also, during the initial recombination screen we noted that only three Zambian, one Botswanan, and two South African *gag* sequences and one Botswanan and three South African *nef* sequences were probable intersubtype recombinants. All of these sequences are previously identified recombinants, but none is a recognized CRF. Our analyses indicated, however, that A2C.ZM.x.ZAM716 and A2C.ZM.89.ZAM174 share nearly identical breakpoints in *gag* and may be examples of a recombinant form currently circulating in Zambia. That only 7.8% of the 64 potentially recombinant southern African *gag* sequences and 3.1% of the 127 potentially recombinant Southern African *nef* sequences are intersubtype recombinants indicates that there is evidence of only limited gene flow from other subtypes or CRFs into the region's overwhelmingly subtype C HIV population. This implies, first, that most all non-C subtypes and all recognized CRFs are either very rare or absent from southern Africa, and second, that there may be a degree of epidemiological isolation between subtype C and the other HIV-1 subtypes that have been identified in the region (such as A, B, and D).

The 373 *gag* and 421 *nef* sequences retained following elimination of potential recombinants were realigned with POA and used to construct neighbor-joining trees (Fig. 1; with transition:transversion ratios estimated from the alignments, pairwise deletion of indels, and 1000 bootstrap and branch length test replicates) with MEGA 3.0.[17]

Both the *gag* and *nef* phylogenetic analyses provide clear evidence of geographical clustering of subtype C isolates in various parts of the world. South American isolates (mostly from Brazil) form single well-supported monophyletic groups (>70% bootstrap and > 95% branch length test support) in both the *gag* and *nef* trees. Also in both trees, the Indian isolates mostly fall into one of two groups. Whereas both groups are well supported (>70% bootstrap and 99% branch length test support) in the *nef* tree, only one is well supported in the

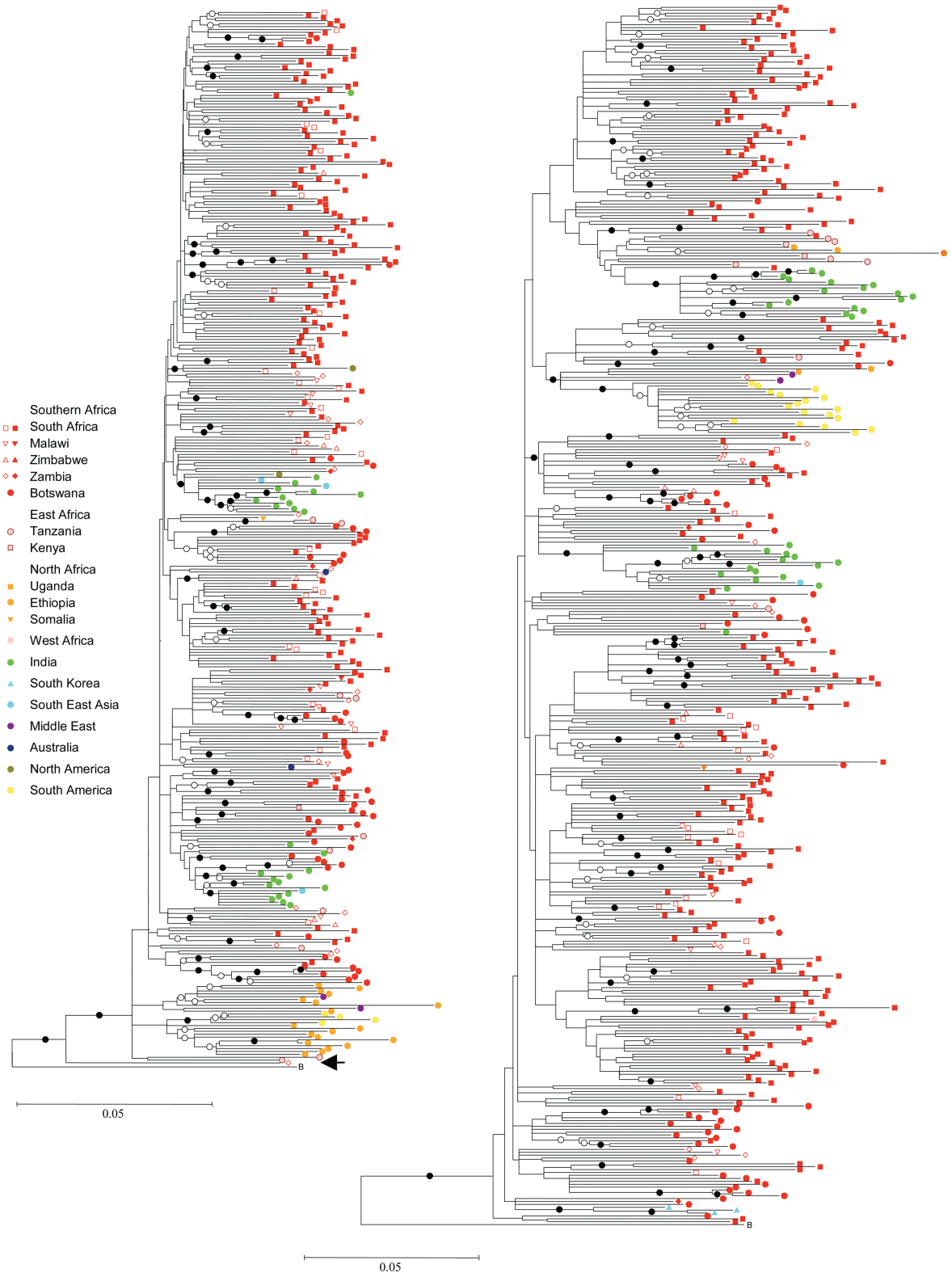

FIG. 1. Neighbor-joining trees indicating the phylogenetic relationships between 382 *gag* (**A**) and 430 *nef* (**B**) HIV-1 subtype C sequences. The trees were constructed using Kimura, 1980 distance matrices with transition:transversion ratios estimated from the data and pairwise deletion of indels (Mega 3.0 Kumar *et al*., 2004). Colored symbols at the ends of branches represent sequences sampled from different regions of the world. Open symbols with a red border represent the 64 *gag* and 44 *nef* sequences sampled from Southern Africa during the present study. Both trees were rooted using the HIV-1 subtype B isolate, HXB2, as an out-group and the significance of their interior branches tested using 1000 bootstrap and interior branch length test replicates. Solid and open black circles indicate branches with greater that 70% and 50% bootstrap support, respectively. All branches with less than 50% bootstrap support and less than 70% interior branch length test support have been collapsed. The arrow indicates a divergent subtype C-like *gag* sequence from Zambia. While the trees were constructed using an additional HIV-1 group N sequence and nine HIV-1 M subtype and subsubtype reference sequences, only the HIV-1 M subtype B reference sequence is shown.

gag tree. The South East Asian subtype C sequences (mostly from China) are clearly most closely related to isolates in one or the other Indian groups, indicating linkage between these epidemics. In contrast, the subtype C epidemic in South Korea (*nef* tree in Fig. 1) has not been obviously founded by viruses from either India or South East Asia. In fact the Korean viruses most closely resemble divergent subtype C isolates from Southern Africa and it is possible that this is their origin.

Within Africa there is also some evidence of geographical clustering of subtype C lineages. For example, the Ethiopian *gag* and *nef* sequences both cluster within two weakly supported (>70% branch length test support) groups. While a close relationship between Brazilian subtype C viruses and those in one of the Ethopian groups is implied in both the *gag* and *nef* trees, there is also reasonable support (>50% bootstrap support) in both trees for an Ethiopian origin of Middle Eastern subtype C viruses.

While there are small, well-supported exclusively South African or Botswanan groups in both trees, these may be the result of sampling bias since the overwhelming majority of southern African sequences have been obtained from these two countries. That sampling bias is responsible for at least some of the apparent clustering of South African sequences is supported by the observation that only 3 of the 30 South African HIVNET028 *gag* sequences and 2 of the 20 South African HIVNET028 *nef* sequences exclusively group with other South African sequences with greater than 50% bootstrap support. Conversely, of the South African HIVNET028 sequences five *gag* sequences and one *nef* sequence are probably (>50% bootstrap support) most closely related to non-South African sequences. Lack of clear geographical clustering of isolates obtained from different southern Africa countries is possibly evidence of the largely unrestricted movement of subtype C genotypes across the entire subcontinent.

The most striking feature of the *gag* and *nef* sequence phylogenies is that the southern African viruses represent the entire breadth of subtype C diversity. The high diversity of the southern African sequences in general, and those sampled in the HIVNET028 study in particular, is evidenced by the fact that in addition to all of them obviously being subtype C sequences, only 11 of 64 *gag* sequences and 11 of 45 *nef* sequences are clearly (>50% bootstrap support) most closely related to other individual subtype C sequences.

One of the HIVNET028 *gag* sequences from Zambia (541–178 indicated by an arrow in Fig. 1A) is actually the most divergent subtype C *gag* sequence yet identified. Sequence 541–178 is also potentially an intrasubtype C recombinant. Five of the six methods used to detect recombination indicated that 541–178 had most likely inherited the sequence between nucleotide positions 959 and 1221 from a Botswanan subtype C virus resembling C.BW.00BW147127 (p-values = 2.97×10^{-3}, 7.26×10^{-5}, 1.14×10^{-3}, 2.13×10^{-2}, 1.66×10^{-4} for the RDP, RECSCAN, MAXCHI, CHIMAERA, and SISCAN methods, respectively). Although the remainder of the 541–178 sequence is clearly subtype C-like (Fig. 1A), it represents a lineage that apparently split from the main subtype C lineage early after the emergence of the ancestral subtype C progenitor.

The high degrees of southern African subtype C *nef* and *gag* gene diversity will increase the difficulty of constructing broadly protective subtype C-specific vaccines based on these genes. The absence of lower diversity, country-specific, southern African subtype C lineages may also seriously hamper the effective use of vaccines specifically targeting smaller subsets of the HIV variants present in the region. On the other hand, there is not much evidence of intersubtype *gag* and *nef* sequence exchange among the region's isolates. This should improve the prospects of the predominantly *nef*- and *gag*-based vaccines set to be tested in southern Africa in that it is unlikely that these will fail due to the emergence of escape recombinants expressing vaccine-targeted epitopes derived from non-subtype C viruses.

Sequences have been deposited in GenBank under the accession numbers DQ792982–DQ793089.

ACKNOWLEDGMENTS

Support for this work was partly through NIH grant N01-AI-45202 as well as the South African AIDS Vaccine Initiative. DPM is supported by the South African National Bioinformatics Network. We would also like to thank the clinical staff at each of the clinical sites for their invaluable help in collecting and shipping samples. The HIVNET028 Study team comprised of: Clive Gray (co-chair), Haynes Sheppard (co-chair), Zambia University Teaching Hospital; Rosemary Musonda, Susan Allen and Michelle Klautzman, University of Alabama; Newton Kumwenda, Malawi College of Medicine; Taha Taha, Johns Hopkins University; Lynn Zijenah, Victoria Aquino, Michael Chirenje, Mike Mbizo, and Ocean Tobaiwa, University of Zimbabwe; David Katzenstein, Stanford University; Glenda Gray, James McIntyre and Armstrong Mafhandu, Perinatal HIV Research Unit, Chris Hani Baragwanath Hospital, Johannesburg; Efthyia Vardas, Mark Colvin, Dudu Msweli, Wendy Dlamini, Gita Ramjee, Salim Abdool Karim, Quarraisha Abdool Karim, Medical Research Council, Durban; Lynn Morris and Natasha Taylor, National Institute for Communicable Diseases; Carolyn Williamson, Helba Bredell and Celia Rademeyer, University of Cape Town; Jorge Flores, Division of AIDS, NIH (DAIDS); Ward Cates, Linda McNeil and Missie Allen, Family Health International.

REFERENCES

1. Bredell H, Williamson C, Sonnenberg P, Martin DJ, and Morris L: Genetic characterization of HIV type 1 from migrant workers in three South African gold mines. AIDS Res Hum Retroviruses 1998;14:677–684.
2. Novitsky V, Smith UR, Gilbert P, *et al.*: Human immunodeficiency virus type 1 subtype C molecular phylogeny: Consensus sequence for an AIDS vaccine design? J Virol 2002;76:5435–5451.
3. Mansky LM and Temin HM: Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 1995;69:5087–5094.
4. Perelson AS, Neumann AU, Markowitz M, Leonard JM, and Ho DD: HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. Science 1996;271:1582–1586.
5. Kelleher AD, Long C, Holmes EC, *et al.*: Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-re-

stricted cytotoxic T lymphocyte responses. J Exp Med 2001;193: 375–386.
6. Wagner R, Leschonsky B, Harrer E, *et al*.: Molecular and functional analysis of a conserved CTL epitope in HIV-1 p24 recognized from a long-term non-progressor: Constraints on immune escape associated with targeting a sequence essential for viral replication. J Immunol 1999;162:3727–3734.
7. Mashishi T, Loubser S, Hide W, *et al*.: Conserved domains of subtype C nef from South African HIV type 1-infected individuals include cytotoxic T lymphocyte epitope-rich regions. AIDS Res Hum Retroviruses 2001;17:1681–1687.
8. Addo MM, Yu XG, Rathod A, *et al*.: Comprehensive epitope analysis of human immunodeficiency virus type 1 (HIV-1)-specific T-cell responses directed against the entire expressed HIV-1 genome demonstrate broadly directed responses, but no correlation to viral load. J Virol 2003;77:2081–2092.
9. Masemola A, Mashishi T, Khoury G, *et al*.: Hierarchical targeting of subtype C human immunodeficiency virus type 1 proteins by $CD8^+$ T cells: Correlation with viral load. J Virol 2004;78:3233–3243.
10. Edwards BH, Bansal A, Sabbaj S, Bakari J, Mulligan MJ, and Goepfert PA: Magnitude of functional CD8+ T-cell responses to the gag protein of human immunodeficiency virus type 1 correlates inversely with viral load in plasma. J Virol 2002;76:2298–2305.
11. Novitsky V, Gilbert P, Peter T, *et al*.: Association between virus-specific T-cell responses and plasma viral load in human immunodeficiency virus type 1 subtype C infection. J Virol 2003;77:882–890.
12. Grasso C and Lee C: Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics 2004;20:546–1556.
13. Felsenstein J: PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 1988;5:164–166.
14. Martin DP, Williamson C, and Posada D: RDP2: Recombination detection and analysis from sequence alignments. Bioinformatics 2005;21:260–262.
15. Schierup MH and Hein J: Consequences of recombination on traditional phylogenetic analysis. Genetics 2000;156:879–891.
16. Posada D and Crandall KA: The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol 2002;54:396–402.
17. Kumar S, Tamura K, and Nei M: MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinf 2004;5:150–163.

Address reprint requests to:
Helba Bredell
Institute of Infectious Disease and Molecular Medicine
Faculty of Health Sciences
University of Cape Town
Observatory
Cape Town, South Africa 7925

E-mail: hbredell@curie.uct.ac.za